Reinforcing Video Reasoning Segmentation to Think Before It Segments

Sitong Gong, Lu Zhang, Yunzhi Zhuge, Xu Jia, Pingping Zhang, Huchuan Lu

IIAU Lab, Dalian University of Technology

Abstract

Video reasoning segmentation (VRS) endeavors to delineate referred objects in videos guided by implicit instructions that encapsulate human intent and temporal logic. Previous approaches leverage large vision language models (LVLMs) to encode object semantics into <SEG> tokens for mask prediction. However, this paradigm suffers from limited interpretability during inference and suboptimal performance due to inadequate spatiotemporal reasoning. Drawing inspiration from seminal breakthroughs in reinforcement learning, we introduce Veason-R1, a specialized LVLM for VRS that emphasizes structured reasoning in segmentation. Veason-R1 is trained through Group Relative Policy Optimization (GRPO) augmented with Chain-of-Thought (CoT) initialization. To begin with, we curate high-quality CoT training data to instill structured reasoning trajectories, bridging video-level semantics and frame-level spatial grounding, yielding the supervised fine-tuned model Veason-SFT. Subsequently, GRPO fine-tuning encourages efficient exploration of the reasoning space by optimizing reasoning chains. To this end, we incorporate a holistic reward mechanism that synergistically enhances spatial alignment and temporal consistency, bolstering keyframe localization and fine-grained grounding. Comprehensive empirical evaluations demonstrate that Veason-R1 achieves state-of-the-art performance on multiple benchmarks, surpassing prior art by significant margins (*e.g.*, +1.3 $\mathcal{J}\&\mathcal{F}$ in ReVOS and +10.0 $\mathcal{J}\&\mathcal{F}$ in ReasonVOS), while exhibiting robustness to hallucinations (+8.8 $\mathcal{R}$). Our code and model weights will be available at Veason-R1.

1. Introduction

Video reasoning segmentation (VRS) (Yan et al. 2024; Bai et al. 2024) aims to produce pixel-wise mask sequences based on language queries encompassing human commonsense and implicit temporal logic. Unlike traditional referring video object segmentation (Seo, Lee, and Han 2020; Wu et al. 2022), which depends on explicit descriptions (*e.g.*, *"the person on a skateboard"*), VRS harnesses world knowledge and temporal modeling in large vision language models (LVLMs) (Liu et al. 2023; Bai et al. 2025a) to perceive complex dynamics with fine granularity. The capability of modeling intricate temporal relations is critical for real-world applications that rely on sequential reasoning to support naunced perception and action, facilitating reliable decisions in domains such as robotic manipulation (Billard and Kragic 2019; Shridhar, Manuelli, and Fox 2022) and autonomous driving (Tian et al. 2024; Xie et al. 2025).

Representative approaches (Yan et al. 2024; Bai et al. 2024; Zheng et al. 2024a; Wei et al. 2024b; Gong et al. 2025) in VRS fields typically employ the LVLM to transform the language query into specialized tokens that serve as semantic embeddings of the referred targets across the video, followed by a mask decoder to produce the corresponding mask trajectories. Despite achieving strong performance, these methods face two key limitations: (i) ***Limited Reasoning and Semantic Alignment.*** Prior approaches inject video-level information into segmentation tokens but lack structured reasoning traces, leading to inherent semantic ambiguity. As shown in Fig. 1 (a), this diminishes efficacy in reasoning-intensive scenarios, such as long videos with temporal occlusions or evolving object interactions, where multi-step inference is essential. (ii) ***Reliance on Large-Scale Training Data.*** Token-based methods demand extensive annotated datasets for LVLM fine-tuning, as associating specialized tokens with image embeddings requires diverse examples for cross-modal alignments and spatiotemporal handling (*e.g.*, motion, occlusions). VISA (Yan et al. 2024) exemplifies this by training on a mixture of image and video segmentation datasets (encompassing 8.8k videos and 214k images), thereby inflating costs and impeding scalability, efficiency, and low-resource generalization.

Recent research (Guo et al. 2025b; Shao et al. 2024) demonstrates that reinforcement learning (RL) fine-tuning elicits structured, interpretable reasoning from large language models (LLMs) at inference time. In particular, Group Relative Policy Optimization (GRPO) enhances in-context reasoning by estimating relative advantages within response groups, enabling critic-free optimization that is both efficient and data-sparse over traditional RL methods. Building on this, several works (Liu et al. 2025a; Wang et al. 2025a; Huang et al. 2025; You and Wu 2025) have applied GRPO to reasoning-based segmentation tasks, using reward functions based on format constraints and IoU metrics to enhance segmentation quality. Meanwhile, others (Feng et al. 2025; Li et al. 2025; Bai et al. 2025b; Zhong et al. 2025) have extended GRPO to video understanding and multi-image grounding, achieving more coherent reasoning and robust grounding in complex visual scenarios. Inspired by these advancements, we propose a RL framework for video reason-

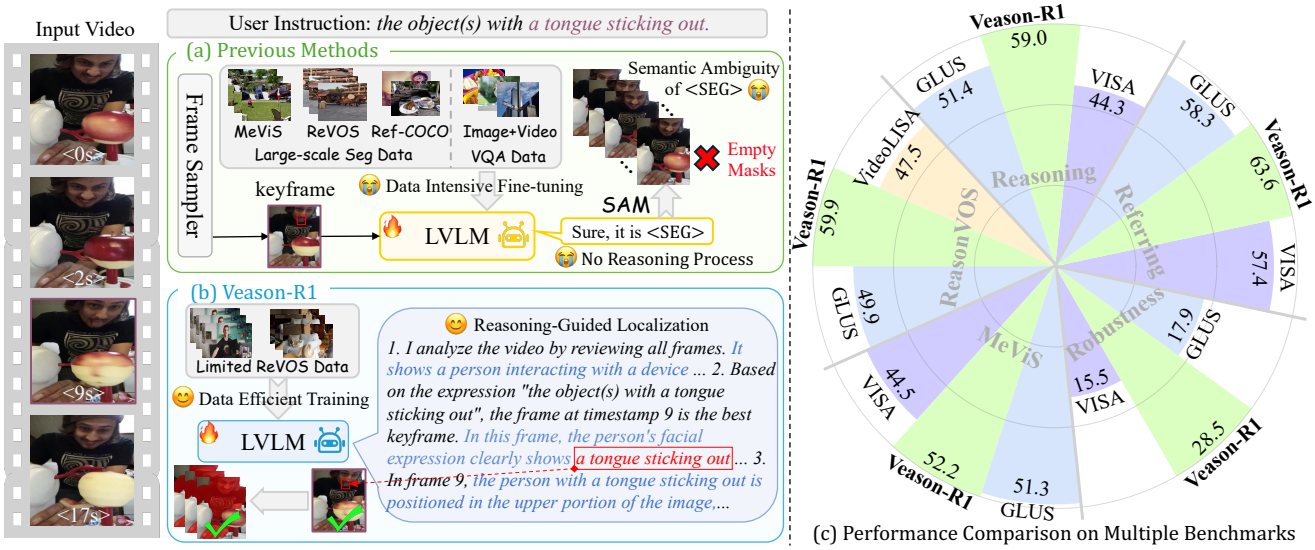


Figure 1: **Comparison of Veason-R1 with Existing Video Reasoning Segmentation (VRS) Approaches**. (a) Conventional methods fine-tune an LVLM using structured data, encoding global video-level information into a single <SEG> token for segmentation guidance; however, this token exhibits weak reasoning capabilities, leading to misaligned masks. (b) In contrast, Veason-R1 introduces a reasoning-guided paradigm enhanced by spatiotemporal reinforcement learning, enabling step-by-step thinking to identify keyframes and produce interpretable segmentation with minimal training data. (c) Veason-R1 achieves SOTA performance on multiple benchmarks, demonstrating superior segmentation accuracy and hallucination robustness.

ing segmentation that integrates Chain-of-Thought (CoT) imitation learning with GRPO-based fine-tuning.

To develop structured reasoning capabilities for identifying critical frames and improving grounding performance in dynamic visual contexts, we first curate a CoT dataset comprising 5.8k annotated samples. This dataset is used to perform supervised fine-tuning (SFT) of the Qwen2.5-VL model (Bai et al. 2025a). Specifically, we employ customized prompt templates to guide Seed1.5-VL (Guo et al. 2025a) in generating CoT reasoning traces, which serve as ground-truth supervision. These reasoning traces guide the model to analyze video content, identify keyframes from referring expressions, and perform accurate grounding within the identified frames. Through this imitation learning process, **Veason-SFT** acquires both keyframe analysis and basic object grounding capabilities. To further hone these abilities, we fine-tune on 10k samples from the ReVOS dataset using GRPO. This stage incorporates a tailored reward policy that synergistically supervises spatial precision and temporal relevance. The reward policy includes a temporal localization reward that evaluates keyframe saliency, a spatial alignment reward that quantifies spatial grounding accuracy, and a unified consistency reward, which integrates SAM2 (Ravi et al. 2024) to fortify the coherence between keyframe selection and spatial grounding. Overall, this two-stage training pipeline yields robust, interpretable reasoning and precise visual grounding in complex videos, as illustrated in Fig. 1 (b). The final **Veason-R1** model attains state-of-the-art performance across multiple VRS benchmarks, exhibiting superior segmentation accuracy and enhanced robustness against hallucinations, as shown in Fig. 1 (c).

Key contributions can be summarized as follows:

- We introduce **Veason-R1**, the first approach to video reasoning segmentation that employs reinforcement learning. Specifically, we leverage GRPO-driven policy optimization, initialized with structured fine-tuning, to jointly enable keyframe identification and spatial grounding using only 10k training samples, a significant reduction compared to the 192k samples required by priors.
- We curate a chain-of-thought (CoT) dataset to equip the model with hierarchical reasoning capabilities, bridging video-level understanding and frame-level object grounding. Furthermore, we design a complementary reward policy during the GRPO stage to enhance temporally coherent reasoning and fine-grained localization.
- Experiments demonstrate that **Veason-R1** consistently obtains superior performance on the ReVOS, ReasonVOS and MeViS benchmarks, validating its effectiveness in logical reasoning and detailed visual comprehension.

2. Related Works

Video Reasoning Segmentation (VRS) is an emerging task requiring explicit multi-modal reasoning to segment target objects in video based on natural language queries (Yan et al. 2024; Bai et al. 2024; Zheng et al. 2024a). Pioneering work such as VISA (Yan et al. 2024) combines keyframe sampling with a large vision language model (LVLM) for temporal reasoning and employs an object tracker for mask propagation. VideoLISA (Bai et al. 2024) introduces a sparse-to-dense sampling strategy and the One-Token-Seg-All paradigm, enabling video-level segmentation via a unified token representation. HyperSeg (Wei et al. 2024a) generalizes the unified reasoning framework through a hybrid entity recognition mechanism, supporting cross-domain seg-

mentation across both image and video inputs. Further advances include VRS-HQ (Gong et al. 2025), enhancing temporal consistency through token fusion and occlusion-aware keyframe selection with SAM2, and Sa2VA (Yuan et al. 2025), integrating SAM2 with LVLM for balanced performance in multiple dense grounding and conversational tasks.

However, such segmentation token-based approaches are limited by the semantic ambiguity of the token representation, often misaligned with the actual objectives and inherently lack interpretability. To address these limitations, we focus on leveraging the GRPO algorithm to enhance the implicit reasoning in VRS, jointly optimizing segmentation performance across spatial and temporal dimensions.

Visual Reinforcement Fine-Tuning. Reinforcement learning (Sutton, Barto et al. 1998) has emerged as a powerful paradigm for optimizing large language models, particularly for enhancing their reasoning capabilities, as demonstrated by ChatGPT-o1(Jaech et al. 2024). Deepseek-R1 (Guo et al. 2025b) introduces group relative policy optimization (GRPO),which leverages verifiable reward signals to estimate relative advantages among responses, thereby substantially improving reasoning. Building on this, GRPO-based fine-tuning has been extended to various multimodal tasks, including image spatial reasoning (Liu et al. 2025c,a; Wang et al. 2025a), video understanding (Feng et al. 2025; Li et al. 2025; Wang et al. 2025b), multi-image grounding (Bai et al. 2025b; Zhang et al. 2025), and visual generation (Fang et al. 2025; Xiao et al. 2025; Xue et al. 2025), showcasing its versatility in challenging multimodal scenarios. Earlier methods such as Seg-Zero (Liu et al. 2025a) and VisionReasoner (Liu et al. 2025b) incorporate segmentation-specific reward designs, leading to improved performance in image-level reasoning segmentation tasks. Inspired by these, Omni-R1 (Zhong et al. 2025) applies GRPO to Ref-AVS and VRS; however, its cascaded dual-system architecture relies on large-scale training data and a pre-trained VRS model. In contrast, our Veason-R1 leverages a single system that simultaneously performs keyframe localization and object grounding using only 10k fine-tuning samples, delivering superior segmentation accuracy in complex video scenarios.

3. Preliminary

Group Relative Policy Optimization (GRPO) (Guo et al. 2025b) is a reinforcement learning algorithm derived from Proximal Policy Optimization (PPO) (Schulman et al. 2017), eliminating the need for a separate value function by using relative rewards from grouped samples. For a given prompt p, GRPO samples G responses $o = \{o_1, ..., o_G\}$ from the policy π_θ. A scalar reward function $r(\cdot)$ evaluates each response, yielding $\{r(o_1), ..., r(o_G)\}$. These rewards are then normalized within the group to compute relative advantages:

$$A_i = \frac{r(o_i) - mean(\{r(o_i)\}_{i=1}^G)}{std(\{r(o_i)\}_{i=1}^G)}, \quad (1)$$

where A_i represents the relative advantage of the i-th candidate response. This encourages the model to favor higher-quality outputs within each sampled group, without relying on absolute reward calibration. To stabilize training and avoid policy collapse, GRPO further introduces a KL-divergence regularization term $D_{KL}(\cdot||\cdot)$ to penalize deviations from a reference policy π_{ref}, resulting in the following optimization objective:

$$\max_{\pi_\theta} \mathbb{E}_{o\sim\pi_{\theta_{\text{old}}}}(p)[\sum_{i=1}^{G}(\frac{\pi_\theta(o_i)}{\pi_{\theta_{\text{old}}(o_i)}})\cdot A_i - \beta D_{KL}(\pi_\theta||\pi_{\text{ref}})], \quad (2)$$

with β serving as a hyperparameter that modulates the intensity of the regularization. This balanced objective encourages the policy model to generate responses with higher relative rewards while preserving proximity to the original policy distribution, supporting reliable convergence in reinforcement learning scenarios.

4. Methodology

Overview

Given a video sequence $\mathcal{V}_I \in \mathbb{R}^{T\times 3\times H\times W}$ consisting of T RGB frames, where H and W denote the height and width of each frame, and a reasoning instruction Q_{txt}, the goal of VRS is to generate a sequence of segmentation masks $\mathcal{M} \in \mathbb{R}^{T\times H\times W}$ using the model.

Unlike prior VRS approaches (Yan et al. 2024; Bai et al. 2024; Zheng et al. 2024a) that embed object semantics into specialized tokens without offering interpretability or transparency, our method explicitly encourages step-by-step reasoning prior to mask generation. We formulate the task as a two-stage process: the model first analyzes the video $\mathcal{V}_I$ in conjunction with the referring expression Q_{txt} to identify a keyframe timestamp T_k in which the referred object is most representative. It then performs spatial grounding by predicting a set of bounding boxes $\mathcal{B}_{T_k} = \{b_i\}_{i=1}^{N_k}$, where N_k is the number of detected instances in the keyframe and each b_i corresponds to a bounding box in (x_1, y_1, x_2, y_2) format. This explicit decomposition enhances both interpretability and spatiotemporal grounding accuracy.

As illustrated in Fig. 2 (a), we adopt a two-stage training paradigm. We initialize our framework with Qwen2.5-VL (Bai et al. 2025a) as the LVLM. In the first stage, we construct a chain-of-thought (CoT) dataset and perform supervised fine-tuning, resulting in the Veason-SFT model. This stage equips the model with hierarchical video reasoning capabilities, enabling it to identify keyframes and perform coarse object localization. In the second stage, we apply GRPO to refine the model's reasoning space, producing the Veason-R1 model with improved spatiotemporal grounding and enhanced reasoning coherence.

CoT-based Supervised Fine-Tuning

Directly applying reinforcement learning (RL) to advanced LVLMs for VRS often leads to unstable training and suboptimal reasoning behavior. This limitation stems from the inability of Qwen2.5-VL to perform structured reasoning under complex temporal dynamics and implicit queries, which hinders effective exploration of reasoning trajectories during RL optimization. To address this challenge, we first construct a high-quality CoT dataset to provide structured cold-start supervision. This initialization pre-equips the model

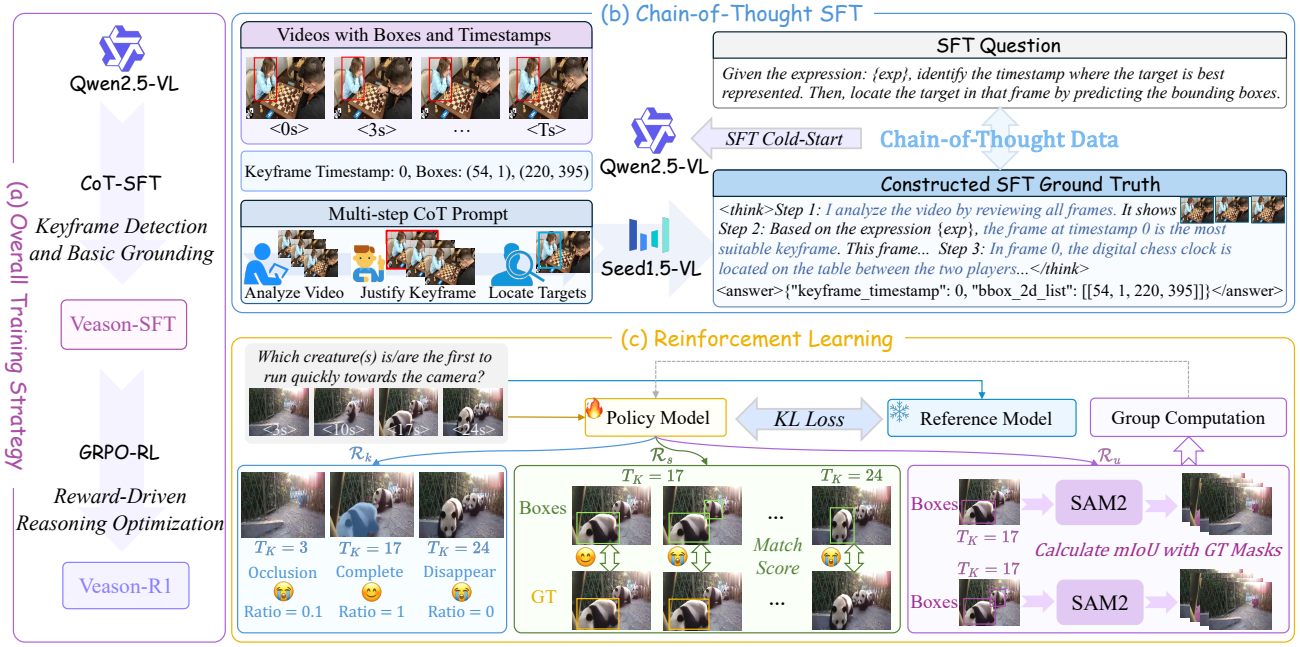


Figure 2: **Overall Training Pipeline of Veason-R1**. (a) Our two-stage training strategy comprises **Veason-SFT** (CoT-based supervised fine-tuning (SFT) for reasoning-aware keyframe selection and basic grounding), followed by **Veason-R1** (reinforcement learning (RL) via GRPO to enhance reasoning fidelity). (b) In the SFT stage, multi-step CoT prompts and pseudo annotations guide Seed1.5-VL to generate structured reasoning traces, which are combined with annotations for LVLM supervision. (c) During RL, the policy model samples candidate keyframes and box predictions. The reward policy jointly reinforces spatial alignment of boxes and temporal consistency in keyframe selection, with KL-regularization ensuring training stability.

with task-specific reasoning priors and guides it to form hierarchical understanding across video-level semantics and frame-level spatial details, which serves as a stable foundation for the subsequent RL fine-tuning.

CoT Data Construction. To construct instructive CoT reasoning traces, we design a data generation pipeline to guide Seed1.5-VL (Guo et al. 2025a) in generating step-by-step reasoning processes, as illustrated in Fig. 2 (b). This approach ensures accurate generated reasoning traces while reducing manual annotation costs. Specifically, we randomly sample one pseudo keyframe from the top-5 frames with the largest target area in each video to prevent the model from overfitting to fixed keyframe indices. Based on this preselected frame, we devise a multi-step CoT prompt that instructs the model to (i) analyze the scene, (ii) substantiate the frame's relevance to the expression, and (iii) localize the referred objects within this frame. Subsequently, we delineate all referred objects using bounding boxes in sampled video frames, and feed the annotations, timestamps, and the predefined CoT prompt into Seed1.5-VL for generating reasoning chains. We encapsulate the final CoT reasoning process within `<think> </think>` tags, and place the selected keyframe and bounding box coordinates inside `<answer> </answer>` tags to serve as the ground truth.

Supervised Fine-Tuning. Leveraging our constructed CoT dataset, we fine-tune the Qwen2.5-VL model to endow it with both keyframe selection and preliminary spatial grounding capabilities. During training, we adopt the auto-regressive next-token prediction paradigm, treating the thinking process and final answer as a unified text sequence and applying cross-entropy loss over the entire generation sequence. Given the limited size of the CoT dataset, we adopt LoRA-based (Hu et al. 2022) fine-tuning to efficiently enhance reasoning capabilities while mitigating overfitting.

Reward-Driven Reasoning Optimization via GRPO

Building upon Veason-SFT's capabilities in keyframe selection and basic grounding, we adopt a GRPO-based RL strategy to further enhance the model's contextual reasoning quality. GRPO uses preference-based rewards to guide the model in generating reasoning chains that more precisely focus on the salient frame and referred targets, enabling it to handle intricate temporal dynamics and generalize across diverse video scenarios. As illustrated in Fig. 2 (c), the policy model explores the action space by generating G candidate outputs per iteration and receives feedback through multiple task-aligned reward signals. Following the GRPO framework, we estimate relative advantages from sampled outputs (*cf* Eq. 1) and update the policy model accordingly (*cf* Eq. 2), with KL regularization for stable optimization.

Given the pivotal role of the reward function in effective policy optimization, we devise a tailored reward mechanism that evaluates reasoning quality from key perspectives, incorporating a temporal localization reward for keyframe saliency, a spatial alignment reward for detection accuracy, and a unified consistency reward for spatiotemporal coher-

ence between selected frames and grounded objects.

Format Compliance Reward. We employ the reward $\mathcal{R}_f$ to incentivize a structured, step-by-step thinking process and enforce strict output formatting: the reasoning trace must be enclosed within `<think> </think>` tags, while the final answer must appear within `<answer> </answer>` tags, formatted as a dictionary containing two mandatory fields, *i.e.*, keyframe_timestamp and bbox_2d_list.

Temporal Localization Reward. The reward $\mathcal{R}_k$ encourages the model to select frames in which the referred object is most visually prominent. Given $\hat{T}$ uniformly sampled video frames $\mathcal{V}_s$, we prepend each image token with their corresponding timestamp (*e.g.*, `<0s>`). For the predicted keyframe timestamp T_k, we compute the reward as the ratio between the mask area of the referred object in frame T_k and the maximum mask area across all sampled frames:

$$\mathcal{R}_k = \frac{\mathcal{S}_{T_k}}{max(\{\mathcal{S}_t\}_{t=1}^{\hat{T}})}, \tag{3}$$

where $\mathcal{S}_t$ represents the foreground area of the frame at t.

Spatial Alignment Reward. The reward $\mathcal{R}_s$ measures the accuracy of object localization in the predicted keyframe. Given the predicted bounding boxes $\mathcal{B}_{T_k} = \{b_i\}_{i=1}^{N_k}$ and the ground-truth boxes $\mathcal{B}_{T_k}^{gt} = \{b_j^{gt}\}_{j=1}^{N_k^{gt}}$ at keyframe T_k, where N_k and N_k^{gt} refer to the number of predicted and ground-truth targets within keyframe, we compute a pairwise IoU-based cost matrix $C \in \mathbb{R}^{N_k \times N_k^{gt}}$ as:

$$C_{i,j} = 1 - IoU(b_i, b_j^{gt}), \tag{4}$$

where $IoU(\cdot)$ denotes the intersection-over-union between two bounding boxes. To support multi-object grounding and enforce one-to-one matching, we apply the Hungarian algorithm to obtain the optimal assignment (i^*, j^*). The final reward is computed as the normalized sum of matched IoUs:

$$\mathcal{R}_s = \frac{1}{\max(N_k, N_k^{gt})} \sum_{(i,j)\in C'} IoU(b_i, b_j^{gt}), \tag{5}$$

where C' denotes the set of matched index pairs derived from the Hungarian algorithm. This formulation effectively promotes precise spatial alignment, even in scenarios involving multiple referred targets.

Unified Consistency Reward. To jointly evaluate the temporal consistency of keyframe selection and the spatial alignment of predicted boxes, we introduce the reward $\mathcal{R}_u$. After performing Hungarian matching on the predicted boxes $\mathcal{B}_{T_k}$ in the selected keyframe, we obtain a set of matched boxes $\mathcal{B}'_{T_k}$ aligned with the ground-truths. These matched boxes, along with the keyframe timestamp, are fed into a frozen SAM2 (Ravi et al. 2024) to generate video-level segmentation masks $\mathcal{M} = \{m_t\}_{t=1}^{\hat{T}}$:

$$\mathcal{M} = \text{SAM2}(\mathcal{B}'_{T_k}, \mathcal{V}_s, T_k) \tag{6}$$

Subsequently, we merge all predicted masks across objects and compute the average IoU with the ground-truth masks $\mathcal{M}^{gt} = \{m_t^{gt}\}_{t=1}^{\hat{T}}$ across sampled frames:

$$\mathcal{R}_u = \frac{1}{\hat{T}} \sum_{t=1}^{\hat{T}} IoU(m_t, m_t^{gt}) \tag{7}$$

The total reward $\mathcal{R}_{total}$ is computed as a weighted sum of the four sub-rewards described above:

$$\mathcal{R}_{total} = \alpha_f \mathcal{R}_f + \alpha_k \mathcal{R}_k + \alpha_s \mathcal{R}_s + \alpha_u \mathcal{R}_u, \tag{8}$$

where the coefficients α_f, α_k, α_s, and α_u are all set to 1.0.

5. Experiments

Datasets and Metrics

We select training samples from the ReVOS dataset, a large-scale VRS benchmark with diverse and complex scenarios, which are crucial for evaluating temporal reasoning and spatial grounding capabilities. In the supervised fine-tuning (SFT) stage, we adapt the Qwen2.5-VL using our curated chain-of-thought dataset tailored for video reasoning segmentation task. During group relative policy optimization (GRPO) training, we sample 10,000 instances proportional to the referring and reasoning subsets in the ReVOS dataset. For evaluation, we rigorously assess Veason-R1 across diverse benchmarks, including the validation set of ReVOS, ReasonVOS (Bai et al. 2024) that features longer videos and more implicit expressions, and MeViS (Ding et al. 2023) that focuses on multi-object and motion-intensive expressions. Consistent with prior works (Yan et al. 2024), we adopt region similarity $\mathcal{J}$, contour accuracy $\mathcal{F}$, and their average $\mathcal{J}\&\mathcal{F}$ as primary metrics. Specifically, $\mathcal{J}$ quantifies the intersection-over-union (IoU) of predicted and ground-truth mask sequence, whereas $\mathcal{F}$ measures the contour-based alignment. In addition, we adopt the robustness score $\mathcal{R}$ to evaluate the model's resistance to hallucination.

Implementation Details

In the supervised fine-tuning (SFT) stage, we use LLaMA-Factory (Zheng et al. 2024b) to fine-tune the Qwen2.5-VL with LoRA (Hu et al. 2022) (rank 8), freezing other parameters. We apply a learning rate of 1×10^{-4}, cosine annealing, and gradient accumulation of 8 steps, training for one epoch on our video-tailored CoT dataset. In the reinforcement learning (RL) stage, we employ the VERL (Sheng et al. 2025) framework with a global batch size of 16, sampling 8 responses per input prompt to facilitate preference optimization, and a learning rate of 1×10^{-6} for one epoch. Experiments are conducted on two NVIDIA A100 GPUs.

Quantitative Comparison

ReVOS. Tab. 1 presents a detailed comparison between Veason-R1 and previous VRS approaches (Yan et al. 2024; Wei et al. 2024b,a; Gong et al. 2025). Notably, despite being fine-tuned on merely 10k samples, Veason-R1-3B attains performance commensurate with the prior state-of-the-art VRS-HQ-13B, while Veason-R1-7B surpasses it by 1.3 in $\mathcal{J}\&\mathcal{F}$. Particularly on the reasoning subset, Veason-R1-7B realizes an improvement of 2.2 in $\mathcal{J}\&\mathcal{F}$, underscoring the efficacy of integrating structured reasoning for guiding precise segmentation under complex temporal dynamics. Furthermore, Veason-R1 manifests a substantially higher robustness score $\mathcal{R}$ than prior methods, indicating that promoting structured reasoning prior to segmentation helps reduce hallucinations and enhance prediction reliability.

Table 1: **Comparative Analysis on the ReVOS Benchmark.** Detailed comparison of Veason-R1's performance against existing approaches on the ReVOS benchmark.

Methods		Referring			Reasoning			Overall			$\mathcal{R}$
		$\mathcal{J}$	$\mathcal{F}$	$\mathcal{J}\&\mathcal{F}$	$\mathcal{J}$	$\mathcal{F}$	$\mathcal{J}\&\mathcal{F}$	$\mathcal{J}$	$\mathcal{F}$	$\mathcal{J}\&\mathcal{F}$	
LMPM (Ding et al. 2023)	[ICCV2023]	29.0	39.1	34.1	13.3	24.3	18.8	21.2	31.7	26.4	3.2
LISA-7B (Lai et al. 2024)	[CVPR2024]	44.3	47.1	45.7	33.8	38.4	36.1	39.1	42.7	40.9	9.3
LISA-13B (Lai et al. 2024)	[CVPR2024]	45.2	47.9	46.6	34.3	39.1	36.7	39.8	43.5	41.6	8.6
VISA-7B (Yan et al. 2024)	[ECCV2024]	49.2	52.6	50.9	40.6	45.4	43.0	44.9	49.0	46.9	15.5
VISA-13B (Yan et al. 2024)	[ECCV2024]	55.6	59.1	57.4	42.0	46.7	44.3	48.8	52.9	50.9	14.5
InstructSeg-3B (Wei et al. 2024b)	[arXiv2025]	54.8	59.2	57.0	49.2	54.7	51.9	52.0	56.9	54.5	-
HyperSeg-3B (Wei et al. 2024a)	[arXiv2025]	56.0	60.9	58.5	50.2	55.8	53.0	53.1	58.4	55.7	-
GLUS-7B (Lin et al. 2025)	[CVPR2025]	56.0	60.7	58.3	48.8	53.9	51.4	52.4	57.3	54.9	17.9
VRS-HQ-7B (Gong et al. 2025)	[CVPR2025]	59.8	64.5	62.1	53.5	58.7	56.1	56.6	61.6	59.1	19.7
VRS-HQ-13B (Gong et al. 2025)	[CVPR2025]	**61.1**	65.5	63.3	54.1	59.4	56.8	57.6	62.5	60.0	18.9
Veason-R1-3B	[Ours]	60.3	65.6	63.0	53.6	60.0	56.8	56.9	62.8	59.9	**28.5**
Veason-R1-7B	[Ours]	60.7	**66.5**	**63.6**	**55.8**	**62.2**	**59.0**	**58.2**	**64.4**	**61.3**	27.0

Table 2: **Benchmark Comparison on the ReasonVOS.**

Methods		$\mathcal{J}$	$\mathcal{F}$	$\mathcal{J}\&\mathcal{F}$
OnlineRefer (Wu et al. 2023)	[ICCV2023]	34.6	42.9	38.7
SgMg (Miao et al. 2023)	[ICCV2023]	33.7	38.7	36.2
LISA-7B (Lai et al. 2024)	[CVPR2024]	29.1	33.1	31.1
VideoLISA-3.8B (Bai et al. 2024)	[NeurIPS2024]	45.1	49.9	47.5
GLUS-7B (Lin et al. 2025)	[CVPR2025]	47.5	52.4	49.9
Veason-R1-3B	[Ours]	51.8	58.5	55.2
Veason-R1-7B	[Ours]	**56.0**	**63.8**	**59.9**

Table 3: **Benchmark Comparison on the MeViS.**

Methods		$\mathcal{J}$	$\mathcal{F}$	$\mathcal{J}\&\mathcal{F}$
LMPM (Ding et al. 2023)	[ICCV2023]	34.2	40.2	37.2
VISA-13B (Yan et al. 2024)	[ECCV2024]	41.8	47.1	44.5
VideoLISA-3.8B (Bai et al. 2024)	[NeurIPS2024]	41.3	47.6	44.4
GLUS-7B (Lin et al. 2025)	[CVPR2025]	48.5	54.2	51.3
VRS-HQ-13B (Gong et al. 2025)	[CVPR2025]	48.0	53.7	50.9
Veason-R1-3B	[Ours]	48.2	54.2	51.2
Veason-R1-7B	[Ours]	**48.4**	**56.0**	**52.2**

ReasonVOS. Tab. 2 compares the segmentation performance of Veason-R1 against prior methods (Bai et al. 2024; Lin et al. 2025) on the ReasonVOS benchmark, which comprises 91 videos with complex scenes (averaging 105 frames each) and 253 long-form queries. These instructions include cases involving causal reasoning (*e.g.*, inferring intent from behavior) and hypothetical scenarios (*e.g.*, conditional or counterfactual prompts). Veason-R1 markedly outperforms GLUS by 8.5 in $\mathcal{J}$, 11.4 in $\mathcal{F}$, and 10.0 in $\mathcal{J}\&\mathcal{F}$, demonstrating its superior ability to handle intricate linguistic reasoning and temporally extended, dynamic video content.

MeViS. In addition to VRS benchmarks, we further assess the generalization ability of Veason-R1 on the MeViS dataset, as illustrated in Tab. 3. In contrast to existing methods that incorporate MeViS into their training data, Veason-R1 is trained solely on 10k samples from ReVOS and evaluated on MeViS in a zero-shot setting. Despite this discrepancy, Veason-R1 still surpasses previous art by 0.9 in $\mathcal{J}\&\mathcal{F}$, demonstrating the robustness and adaptability of our "thinking before segmenting" paradigm in tackling motion-centric and referring-based queries without task-specific tuning.

Qualitative Comparison

Fig. 3 qualitatively compares Veason-R1-3B with VISA-7B in challenging scenarios involving occlusions and temporally grounded expressions. For the left example, where the target emerges only in the final frames, Veason-R1 uses temporal reasoning to pinpoint the frame where the warthog is most prominent and accurately localizes it, whereas VISA fails to detect and outputs an empty mask sequence. In the right case, Veason-R1 accurately interprets the visual scene and query to locate the girl "situated by the window" at the video end. These results underscore Veason-R1's strength in handling temporally sensitive and spatially complex instructions via explicit step-by-step reasoning. Additional qualitative results are available in the *Supplementary Materials*.

Table 4: **Ablation Study of Reward Functions.** Combining all reward components achieves the highest performance.

Reward	Referring			Reasoning		
	$\mathcal{J}$	$\mathcal{F}$	$\mathcal{J}\&\mathcal{F}$	$\mathcal{J}$	$\mathcal{F}$	$\mathcal{J}\&\mathcal{F}$
w/o $\mathcal{R}_k$	58.8	64.4	61.6	51.9	57.9	54.9
w/o $\mathcal{R}_s$	58.1	63.9	61.0	50.3	57.5	53.9
w/o $\mathcal{R}_u$	59.6	65.3	62.4	53.0	59.3	56.1
Ours	**60.3**	**65.6**	**63.0**	**53.6**	**60.0**	**56.8**

Table 5: **Ablation Study of Training Strategy.** The integration of CoT-SFT and GRPO achieves the best result.

Strategy	Referring			Reasoning		
	$\mathcal{J}$	$\mathcal{F}$	$\mathcal{J}\&\mathcal{F}$	$\mathcal{J}$	$\mathcal{F}$	$\mathcal{J}\&\mathcal{F}$
Qwen2.5-VL	19.1	22.8	20.9	17.3	21.7	19.5
SFT	36.4	40.2	38.3	31.0	35.3	33.1
CoT-SFT	48.1	54.1	51.1	37.9	45.3	41.6
Pure GRPO	57.0	63.0	60.0	50.8	57.6	54.2
SFT+GRPO	58.9	64.4	61.6	52.8	59.2	56.0
CoT-SFT+GRPO	**60.3**	**65.6**	**63.0**	**53.6**	**60.0**	**56.8**

Ablation Study

We perform extensive ablation study based on Qwen2.5-VL-3B to assess the efficacy of the proposed reward policy, training strategy, and keyframe-first grounding strategy.

Reward Policy. The results of ablating components of the reward functions are presented in Tab. 4. Removing the $\mathcal{R}_s$ induces the most substantial performance decline (2.0 and 2.9 in $\mathcal{J}\&\mathcal{F}$ on referring and reasoning subsets), underscoring the critical role of frame-level spatial alignment in

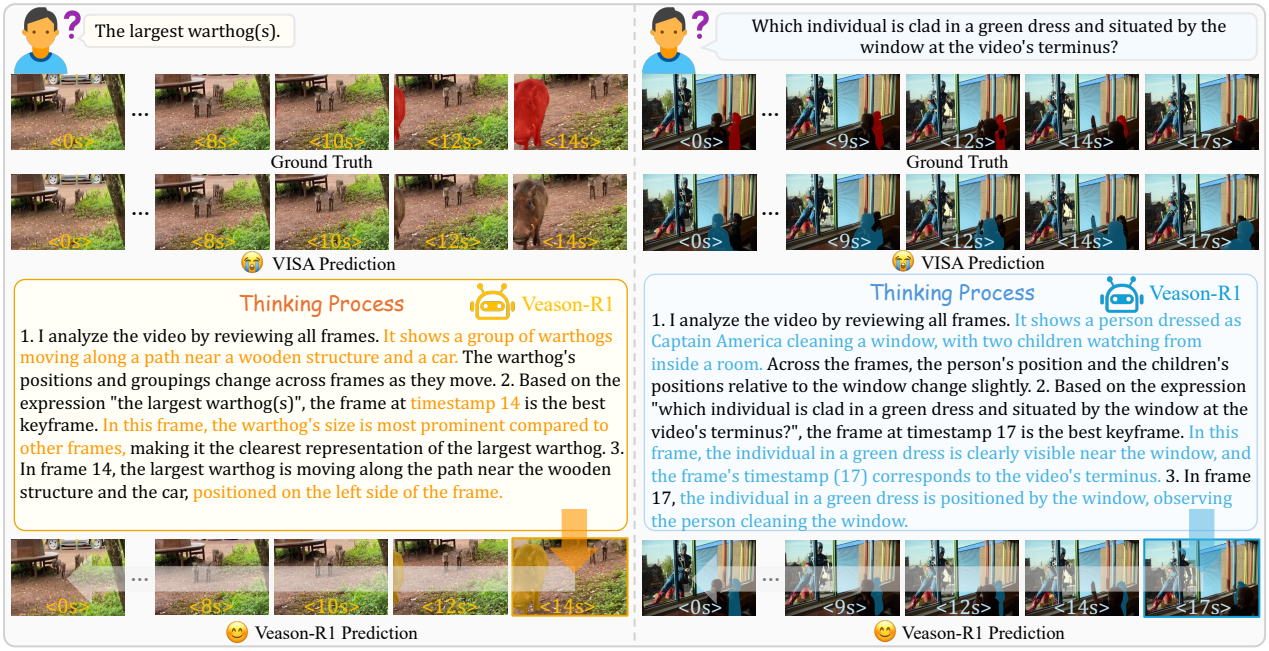


Figure 3: **Qualitative Comparison between Veason-R1-3B with VISA-7B**. In challenging scenarios where the target disappears for a long duration (left) or the expression is highly coupled with temporal context (right), Veason-R1 generates a clear and interpretable reasoning process, accurately selects the keyframe, and reliably segments the semantically referred object.

Table 6: **Ablation Study on Joint Keyframe-grounding Training.** Joint training of keyframe selection and spatial grounding results in best performance.

Strategy	Referring			Reasoning		
	$\mathcal{J}$	$\mathcal{F}$	$\mathcal{J}\&\mathcal{F}$	$\mathcal{J}$	$\mathcal{F}$	$\mathcal{J}\&\mathcal{F}$
Grounding-only	55.8	60.9	58.3	48.2	54.4	51.3
SFT-keyframe	49.7	59.4	54.6	46.5	55.6	51.0
Ours	**60.3**	**65.6**	**63.0**	**53.6**	**60.0**	**56.8**

enhancing the model's localization capability. Furthermore, omitting $\mathcal{R}_u$ yields performance degradations of 0.6 and 0.7 in $\mathcal{J}\&\mathcal{F}$ for the two subsets, demonstrating the effectiveness of jointly optimizing keyframe selection and spatial localization to maintain their temporal coherence.

Training Strategy. Tab. 5 presents the performance comparison across various fine-tuning strategies. We first assess the impact of incorporating the chain-of-thought (CoT) process during the SFT stage. Removing CoT reasoning leads to substantial performance drops of 12.8 and 8.5 in $\mathcal{J}\&\mathcal{F}$ for the referring and reasoning subsets, respectively. Even after subsequent GRPO fine-tuning, the model still lags behind Veason-R1 by 1.4 and 0.8 in $\mathcal{J}\&\mathcal{F}$, underscoring the effectiveness of explicitly supervising the reasoning process for improving both reasoning quality and grounding accuracy. Next, we evaluate the effect of training with SFT and GRPO separately. We find that applying GRPO alone brings notable gains, outperforming CoT-SFT by 8.9 and 12.6 in $\mathcal{J}\&\mathcal{F}$. However, combining CoT-SFT with GRPO yields the best overall results, surpassing GRPO-only by 3.0 (referring) and 2.6 (reasoning) in $\mathcal{J}\&\mathcal{F}$, indicating the complementary benefits of initializing the model with structured reasoning chains before preference optimization.

Keyframe Selection and Grounding. We investigate the impact of jointly modeling keyframe selection and grounding as shown in Tab. 6. Grounding-only indicates directly applying GRPO with the first target-containing frame fixed as the keyframe, while SFT-keyframe denotes training the model exclusively for keyframe detection during the SFT stage. Directly applying GRPO with a fixed keyframe significantly decreases performance by 4.7 and 5.5 in $\mathcal{J}\&\mathcal{F}$ for the referring and reasoning subsets, highlighting the critical need for accurate temporal localization prior to spatial grounding. Furthermore, training the model to perform keyframe selection alone during SFT, while ignoring spatial grounding results in performance degradations of 8.4 and 5.8 in $\mathcal{J}\&\mathcal{F}$, underscoring the importance of jointly modeling temporal and spatial reasoning for optimal segmentation.

6. Conclusion

We present Veason-R1, a reinforcement learning framework for VRS that explicitly models interpretable reasoning trajectories by decomposing the task into keyframe selection and object localization. A two-stage training strategy is employed: first, a CoT-based SFT procedure instills hierarchical reasoning capabilities. Then, GRPO-based RL further refines reasoning and grounding behavior, guided by the reward policy that comprehensively captures spatial alignment, keyframe saliency, and temporal consistency. Veason-R1 achieves strong performance on multiple benchmarks, underscoring the effectiveness of structured pre-segmentation reasoning in enhancing multi-modal understanding of semantic and motion cues over time.

References

Bai, S.; Chen, K.; Liu, X.; Wang, J.; Ge, W.; Song, S.; Dang, K.; Wang, P.; Wang, S.; Tang, J.; et al. 2025a. Qwen2. 5-vl technical report. *arXiv preprint arXiv:2502.13923*.

Bai, S.; Li, M.; Liu, Y.; Tang, J.; Zhang, H.; Sun, L.; Chu, X.; and Tang, Y. 2025b. Univg-r1: Reasoning guided universal visual grounding with reinforcement learning. *arXiv preprint arXiv:2505.14231*.

Bai, Z.; He, T.; Mei, H.; Wang, P.; Gao, Z.; Chen, J.; Zhang, Z.; and Shou, M. Z. 2024. One token to seg them all: Language instructed reasoning segmentation in videos. *Adv. Neural Inform. Process. Syst.*, 37: 6833–6859.

Billard, A.; and Kragic, D. 2019. Trends and challenges in robot manipulation. *Science*, 364(6446): eaat8414.

Ding, H.; Liu, C.; He, S.; Jiang, X.; and Loy, C. C. 2023. MeViS: A large-scale benchmark for video segmentation with motion expressions. In *Int. Conf. Comput. Vis.*, 2694–2703.

Fang, R.; Duan, C.; Wang, K.; Huang, L.; Li, H.; Yan, S.; Tian, H.; Zeng, X.; Zhao, R.; Dai, J.; et al. 2025. Got: Unleashing reasoning capability of multimodal large language model for visual generation and editing. *arXiv preprint arXiv:2503.10639*.

Feng, K.; Gong, K.; Li, B.; Guo, Z.; Wang, Y.; Peng, T.; Wu, J.; Zhang, X.; Wang, B.; and Yue, X. 2025. Video-r1: Reinforcing video reasoning in mllms. *arXiv preprint arXiv:2503.21776*.

Gong, S.; Zhuge, Y.; Zhang, L.; Yang, Z.; Zhang, P.; and Lu, H. 2025. The Devil is in Temporal Token: High Quality Video Reasoning Segmentation. *arXiv preprint arXiv:2501.08549*.

Guo, D.; Wu, F.; Zhu, F.; Leng, F.; Shi, G.; Chen, H.; Fan, H.; Wang, J.; Jiang, J.; Wang, J.; et al. 2025a. Seed1. 5-vl technical report. *arXiv preprint arXiv:2505.07062*.

Guo, D.; Yang, D.; Zhang, H.; Song, J.; Zhang, R.; Xu, R.; Zhu, Q.; Ma, S.; Wang, P.; Bi, X.; et al. 2025b. Deepseek-r1: Incentivizing reasoning capability in llms via reinforcement learning. *arXiv preprint arXiv:2501.12948*.

Hu, E. J.; Shen, Y.; Wallis, P.; Allen-Zhu, Z.; Li, Y.; Wang, S.; Wang, L.; Chen, W.; et al. 2022. Lora: Low-rank adaptation of large language models. *Int. Conf. Learn. Represent.*, 1(2): 3.

Huang, J.; Xu, Z.; Zhou, J.; Liu, T.; Xiao, Y.; Ou, M.; Ji, B.; Li, X.; and Yuan, K. 2025. SAM-R1: Leveraging SAM for Reward Feedback in Multimodal Segmentation via Reinforcement Learning. *arXiv preprint arXiv:2505.22596*.

Jaech, A.; Kalai, A.; Lerer, A.; Richardson, A.; El-Kishky, A.; Low, A.; Helyar, A.; Madry, A.; Beutel, A.; Carney, A.; et al. 2024. Openai o1 system card. *arXiv preprint arXiv:2412.16720*.

Lai, X.; Tian, Z.; Chen, Y.; Li, Y.; Yuan, Y.; Liu, S.; and Jia, J. 2024. Lisa: Reasoning segmentation via large language model. In *IEEE Conf. Comput. Vis. Pattern Recog.*, 9579–9589.

Li, X.; Yan, Z.; Meng, D.; Dong, L.; Zeng, X.; He, Y.; Wang, Y.; Qiao, Y.; Wang, Y.; and Wang, L. 2025. Videochat-r1: Enhancing spatio-temporal perception via reinforcement fine-tuning. *arXiv preprint arXiv:2504.06958*.

Lin, L.; Yu, X.; Pang, Z.; and Wang, Y.-X. 2025. Glus: Global-local reasoning unified into a single large language model for video segmentation. In *IEEE Conf. Comput. Vis. Pattern Recog.*, 8658–8667.

Liu, H.; Li, C.; Wu, Q.; and Lee, Y. J. 2023. Visual instruction tuning. *Adv. Neural Inform. Process. Syst.*, 36: 34892–34916.

Liu, Y.; Peng, B.; Zhong, Z.; Yue, Z.; Lu, F.; Yu, B.; and Jia, J. 2025a. Seg-zero: Reasoning-chain guided segmentation via cognitive reinforcement. *arXiv preprint arXiv:2503.06520*.

Liu, Y.; Qu, T.; Zhong, Z.; Peng, B.; Liu, S.; Yu, B.; and Jia, J. 2025b. VisionReasoner: Unified Visual Perception and Reasoning via Reinforcement Learning. *arXiv preprint arXiv:2505.12081*.

Liu, Z.; Sun, Z.; Zang, Y.; Dong, X.; Cao, Y.; Duan, H.; Lin, D.; and Wang, J. 2025c. Visual-rft: Visual reinforcement fine-tuning. *arXiv preprint arXiv:2503.01785*.

Miao, B.; Bennamoun, M.; Gao, Y.; and Mian, A. 2023. Spectrum-guided multi-granularity referring video object segmentation. In *Int. Conf. Comput. Vis.*, 920–930.

Ravi, N.; Gabeur, V.; Hu, Y.-T.; Hu, R.; Ryali, C.; Ma, T.; Khedr, H.; Rädle, R.; Rolland, C.; Gustafson, L.; et al. 2024. Sam 2: Segment anything in images and videos. *arXiv preprint arXiv:2408.00714*.

Schulman, J.; Wolski, F.; Dhariwal, P.; Radford, A.; and Klimov, O. 2017. Proximal policy optimization algorithms. *arXiv preprint arXiv:1707.06347*.

Seo, S.; Lee, J.-Y.; and Han, B. 2020. Urvos: Unified referring video object segmentation network with a large-scale benchmark. In *Eur. Conf. Comput. Vis.*, 208–223. Springer.

Shao, Z.; Wang, P.; Zhu, Q.; Xu, R.; Song, J.; Bi, X.; Zhang, H.; Zhang, M.; Li, Y.; Wu, Y.; et al. 2024. Deepseekmath: Pushing the limits of mathematical reasoning in open language models. *arXiv preprint arXiv:2402.03300*.

Sheng, G.; Zhang, C.; Ye, Z.; Wu, X.; Zhang, W.; Zhang, R.; Peng, Y.; Lin, H.; and Wu, C. 2025. Hybridflow: A flexible and efficient rlhf framework. In *Proceedings of the Twentieth European Conference on Computer Systems*, 1279–1297.

Shridhar, M.; Manuelli, L.; and Fox, D. 2022. Cliport: What and where pathways for robotic manipulation. In *Conference on robot learning*, 894–906. PMLR.

Sutton, R. S.; Barto, A. G.; et al. 1998. *Reinforcement learning: An introduction*, volume 1. MIT press Cambridge.

Tian, X.; Gu, J.; Li, B.; Liu, Y.; Wang, Y.; Zhao, Z.; Zhan, K.; Jia, P.; Lang, X.; and Zhao, H. 2024. Drivevlm: The convergence of autonomous driving and large vision-language models. *arXiv preprint arXiv:2402.12289*.

Wang, S.; Fang, G.; Kong, L.; Li, X.; Xu, J.; Yang, S.; Li, Q.; Zhu, J.; and Wang, X. 2025a. PixelThink: Towards Efficient Chain-of-Pixel Reasoning. *arXiv preprint arXiv:2505.23727*.

Wang, Y.; Wang, Z.; Xu, B.; Du, Y.; Lin, K.; Xiao, Z.; Yue, Z.; Ju, J.; Zhang, L.; Yang, D.; et al. 2025b. Time-R1: Post-Training Large Vision Language Model for Temporal Video Grounding. *arXiv preprint arXiv:2503.13377*.

Wei, C.; Zhong, Y.; Tan, H.; Liu, Y.; Zhao, Z.; Hu, J.; and Yang, Y. 2024a. HyperSeg: Towards Universal Visual Segmentation with Large Language Model. *arXiv preprint arXiv:2411.17606*.

Wei, C.; Zhong, Y.; Tan, H.; Zeng, Y.; Liu, Y.; Zhao, Z.; and Yang, Y. 2024b. InstructSeg: Unifying Instructed Visual Segmentation with Multi-modal Large Language Models. *arXiv preprint arXiv:2412.14006*.

Wu, D.; Wang, T.; Zhang, Y.; Zhang, X.; and Shen, J. 2023. Onlinerefer: A simple online baseline for referring video object segmentation. In *Int. Conf. Comput. Vis.*, 2761–2770.

Wu, J.; Jiang, Y.; Sun, P.; Yuan, Z.; and Luo, P. 2022. Language as queries for referring video object segmentation. In *IEEE Conf. Comput. Vis. Pattern Recog.*, 4974–4984.

Xiao, Y.; Song, L.; Chen, Y.; Luo, Y.; Chen, Y.; Gan, Y.; Huang, W.; Li, X.; Qi, X.; and Shan, Y. 2025. Mindomni: Unleashing reasoning generation in vision language models with rgpo. *arXiv preprint arXiv:2505.13031*.

Xie, S.; Kong, L.; Dong, Y.; Sima, C.; Zhang, W.; Chen, Q. A.; Liu, Z.; and Pan, L. 2025. Are VLMs Ready for Autonomous Driving? An Empirical Study from the Reliability, Data, and Metric Perspectives. *arXiv preprint arXiv:2501.04003*.

Xue, Z.; Wu, J.; Gao, Y.; Kong, F.; Zhu, L.; Chen, M.; Liu, Z.; Liu, W.; Guo, Q.; Huang, W.; et al. 2025. DanceGRPO: Unleashing GRPO on Visual Generation. *arXiv preprint arXiv:2505.07818*.

Yan, C.; Wang, H.; Yan, S.; Jiang, X.; Hu, Y.; Kang, G.; Xie, W.; and Gavves, E. 2024. Visa: Reasoning video object segmentation via large language models. In *Eur. Conf. Comput. Vis.*, 98–115. Springer.

You, Z.; and Wu, Z. 2025. Seg-R1: Segmentation Can Be Surprisingly Simple with Reinforcement Learning. *arXiv preprint arXiv:2506.22624*.

Yuan, H.; Li, X.; Zhang, T.; Huang, Z.; Xu, S.; Ji, S.; Tong, Y.; Qi, L.; Feng, J.; and Yang, M.-H. 2025. Sa2va: Marrying sam2 with llava for dense grounded understanding of images and videos. *arXiv preprint arXiv:2501.04001*.

Zhang, B.; Li, H.; Zhang, T.; Yan, C.; Cai, J.; Jiang, X.; and Hao, Y. 2025. Improving the Reasoning of Multi-Image Grounding in MLLMs via Reinforcement Learning. *arXiv preprint arXiv:2507.00748*.

Zheng, R.; Qi, L.; Chen, X.; Wang, Y.; Wang, K.; Qiao, Y.; and Zhao, H. 2024a. ViLLa: Video Reasoning Segmentation with Large Language Model. *arXiv preprint arXiv:2407.14500*.

Zheng, Y.; Zhang, R.; Zhang, J.; Ye, Y.; Luo, Z.; Feng, Z.; and Ma, Y. 2024b. Llamafactory: Unified efficient fine-tuning of 100+ language models. *arXiv preprint arXiv:2403.13372*.

Zhong, H.; Zhu, M.; Du, Z.; Huang, Z.; Zhao, C.; Liu, M.; Wang, W.; Chen, H.; and Shen, C. 2025. Omni-R1: Reinforcement Learning for Omnimodal Reasoning via Two-System Collaboration. *arXiv preprint arXiv:2505.20256*.

Appendix

In this supplementary material, we provide additional experimental details and qualitative analyses for Veason-R1. To begin with, we elaborate on the construction of the chain-of-thought training data (Section A). We then provide further implementation details to enhance reproducibility (Section B). Next, we analyze the training curves to illustrate the model's optimization dynamics and convergence behavior (Section C). This is followed by a discussion of representative failure cases to identify potential areas for improvement (Section D). Finally, we include additional qualitative visualizations to further demonstrate the effectiveness of Veason-R1 (Section E).

A. More Details of CoT Training Data Construction

Fig. 4 illustrates the full prompt that we use as input to Seed1.5-VL (Guo et al. 2025a) for chain-of-thought generation. The prompt is presented to the model along with a sequence of uniformly sampled video frames and their corresponding timestamps, as well as the predefined object description and a selected keyframe. The prompt guides the model to (i) briefly summarize the overall video content, (ii) justify why the selected keyframe best matches the target expression, and (iii) describe the target's actions and location within the given frame. This structured prompting encourages the model to produce clear and interpretable reasoning based on the visual and contextual information. We observe that Seed1.5-VL tends to closely mimic the style of human-provided exemplars. To promote diversity in the generated CoT responses and avoid overly rigid reasoning patterns, we design three distinct CoT templates—each with slightly different phrasing and logical flow. For each sample, we randomly select one template during prompting, encouraging the model to explore different reasoning trajectories while maintaining consistency with the underlying task structure.

B. More Implementation Details

We adopt the same question template for both the SFT and GRPO stages, as illustrated in Fig. 5. To improve keyframe grounding, we prepend real timestamps (e.g., `<0s>, <3s>`) before each sampled frame's image tokens, enabling the model to reason with accurate temporal alignment. During the GRPO stage, we set the maximum response length to 1024 tokens, the maximum gradient norm for clipping to 1.0, and the coefficient for the KL divergence regularization loss to 5e-3. All input frames in both training stages are resized to a resolution of 560×560 pixels. During inference, we prompt Veason-R1 to generate the keyframe timestamp and the bounding boxes of the referred objects. These predictions are then passed to SAM2 (Ravi et al. 2024), which performs segmentation and mask prop-

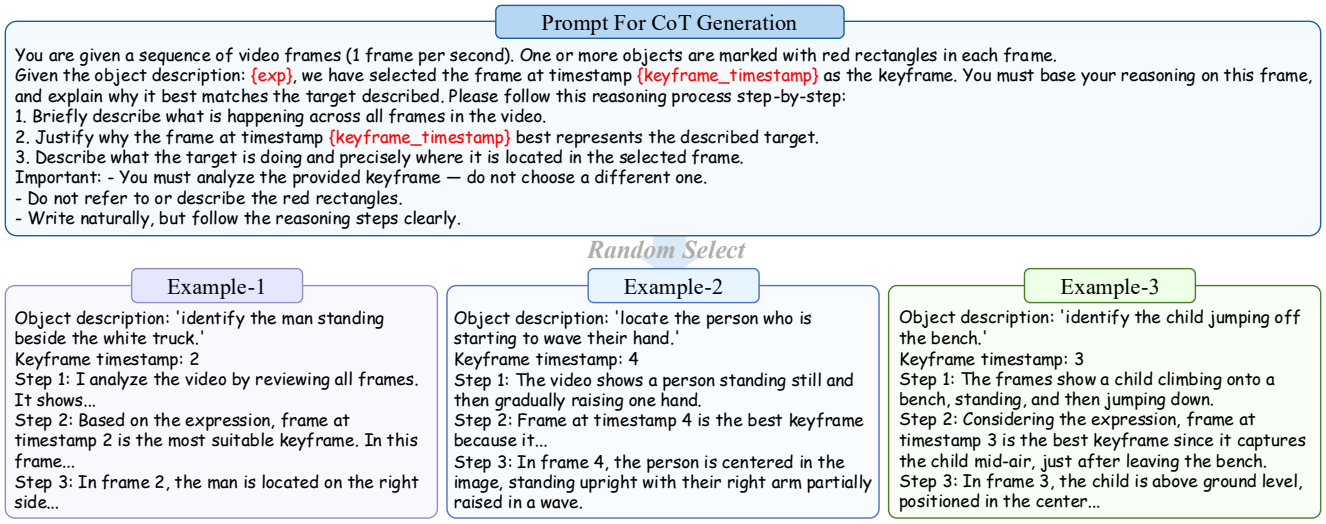


Figure 4: Prompt template for CoT data generation.

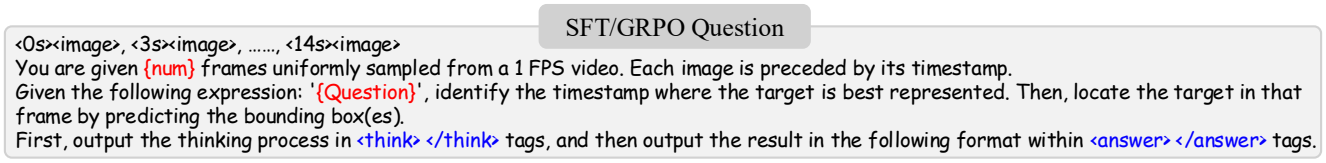


Figure 5: Question template of SFT and GRPO training stage.

agation across the entire video sequence, yielding the final predicted mask sequence.

C. Training Curve Analysis

We visualize the training dynamics of Veason-R1 in Fig. 6. We monitor three key metrics during training: total reward score $\mathcal{R}_{total}$, response length, and the value of KL loss.

As shown in Fig. 6 (a), the model initialized via CoT-based supervision starts with a reward score above 3.0, indicating a reasonable ability to generate structured outputs and locate keyframes. The consistent upward trend in reward reflects effective policy refinement and better alignment with the designed reward function.

It can be observed in Fig. 6 (b) that the average response length initially decreases, suggesting that the model rapidly learns to eliminate redundant tokens and generate concise reasoning. In the later phase, the length slightly increases and then stabilizes, indicating an adjustment towards producing responses that balance brevity and informativeness.

As depicted in Fig. 6 (c), the KL divergence gradually increases as the model departs from the initial supervised policy to explore more reward-driven behaviors. It eventually stabilizes, indicating convergence to a consistent and reward-aligned policy under the regularization constraint.

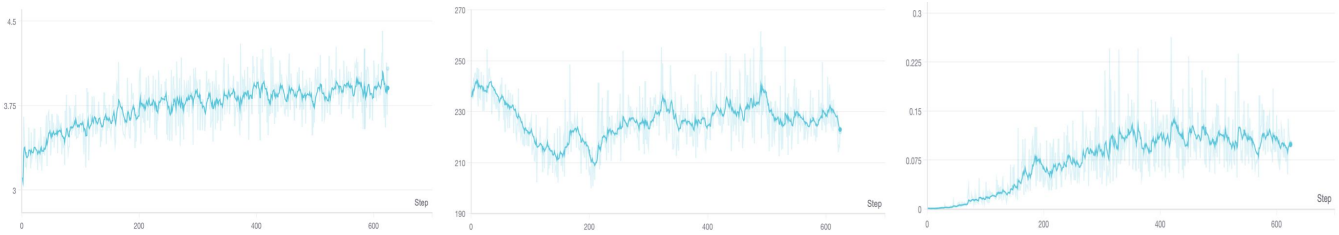
D. Failure Case Analysis

Fig. 7 presents some failure cases of Veason-R1 on the ReVOS dataset. In the left example, the model generates a correct reasoning trace, accurately identifying the tiger closest to the camera and describing its relation to the background tigers. However, the predicted masks fail to align with the described target, revealing a disconnect between reasoning and visual grounding. This suggests that Veason-R1 struggles to translate reasoning into precise spatial localization, especially in cluttered scenes with visually similar objects. In the right example, Veason-R1 correctly identifies the keyframe showing clear motion cues but ultimately selects the slower ship. The reasoning trace overlooks the constraint "fastest" and focuses solely on movement direction. This indicates the model's limited sensitivity to speed and motion-related concepts, revealing weaknesses in capturing temporal dynamics or comparing relative movements. These limitations suggest the importance of incorporating data that better emphasizes motion perception in future work.

E. More Visualization Results

We present additional visual comparisons of Veason-R1-3B with VISA-7B (Yan et al. 2024) and VideoLISA-3.8B (Bai et al. 2024) to underscore its robust reasoning capabilities and enhanced fine-grained grounding performance.

The qualitative comparison of Veason-R1 with VISA is shown in Fig. 8. In the left example, a herd of elephants



(a) Average Reward Score (b) Average Response Length (c) KL Loss

Figure 6: GRPO training curves.

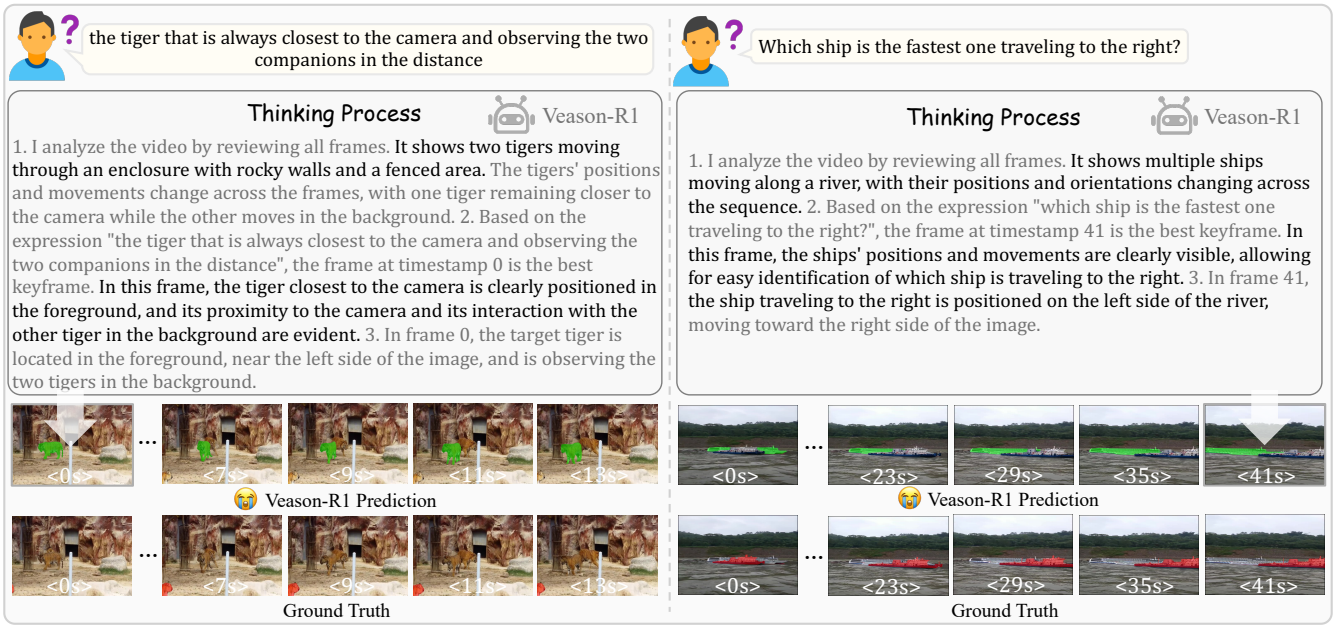


Figure 7: Visualization of failure cases on the ReVOS dataset. Veason-R1 exhibits inconsistencies between its reasoning traces and final segmentation outputs, and shows limited sensitivity to motion-related concepts such as object speed.

partially occludes the vehicles crossing the road. Veason-R1 successfully identifies the timestamp where both vehicles are most clearly visible and segments them, whereas VISA mistakenly highlights background trees. This reflects Veason-R1's robustness to occlusions and its focus on task-relevant semantics. The right case indicates that Veason-R1 correctly segments the slippers in contact with the puppy's feet, demonstrating strong multi-object localization and spatial reasoning. VISA, by contrast, produces incorrect masks, revealing limitations in handling fine-grained interactions.

Fig. 9 visualizes the segmentation maps of Veason-R1 and VideoLISA on the ReasonVOS dataset. In the left example, Veason-R1 accurately segments the dancing bride in response to a poetic query involving abstract movement descriptions, whereas VideoLISA mistakenly segments the groom. This highlights Veason-R1's superior understanding of implicit and abstract language. In the right example, under a long-duration and high-complexity scenario, Veason-R1 successfully localizes the man riding the bicycle in response to a causal reasoning query, while VideoLISA yields fragmented and inconsistent predictions, demonstrating Veason-R1's stronger capacity for temporal reasoning and grounding in complex visual contexts.

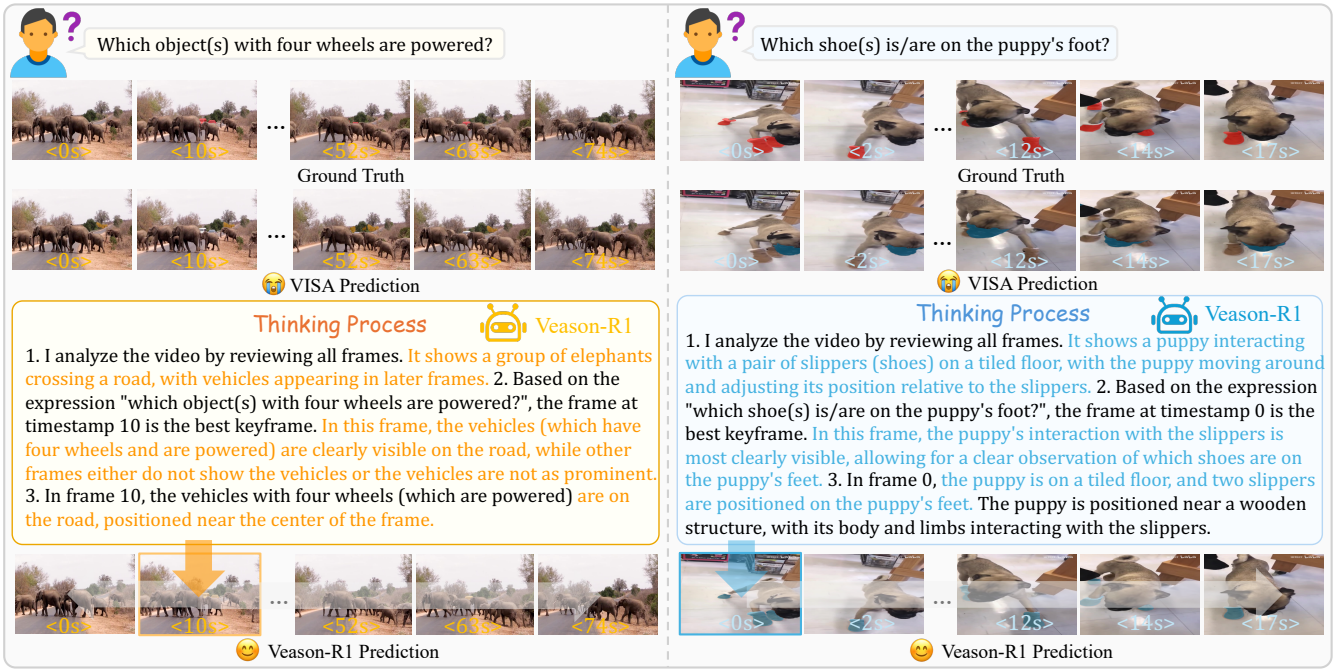


Figure 8: More visualization on the ReVOS dataset.

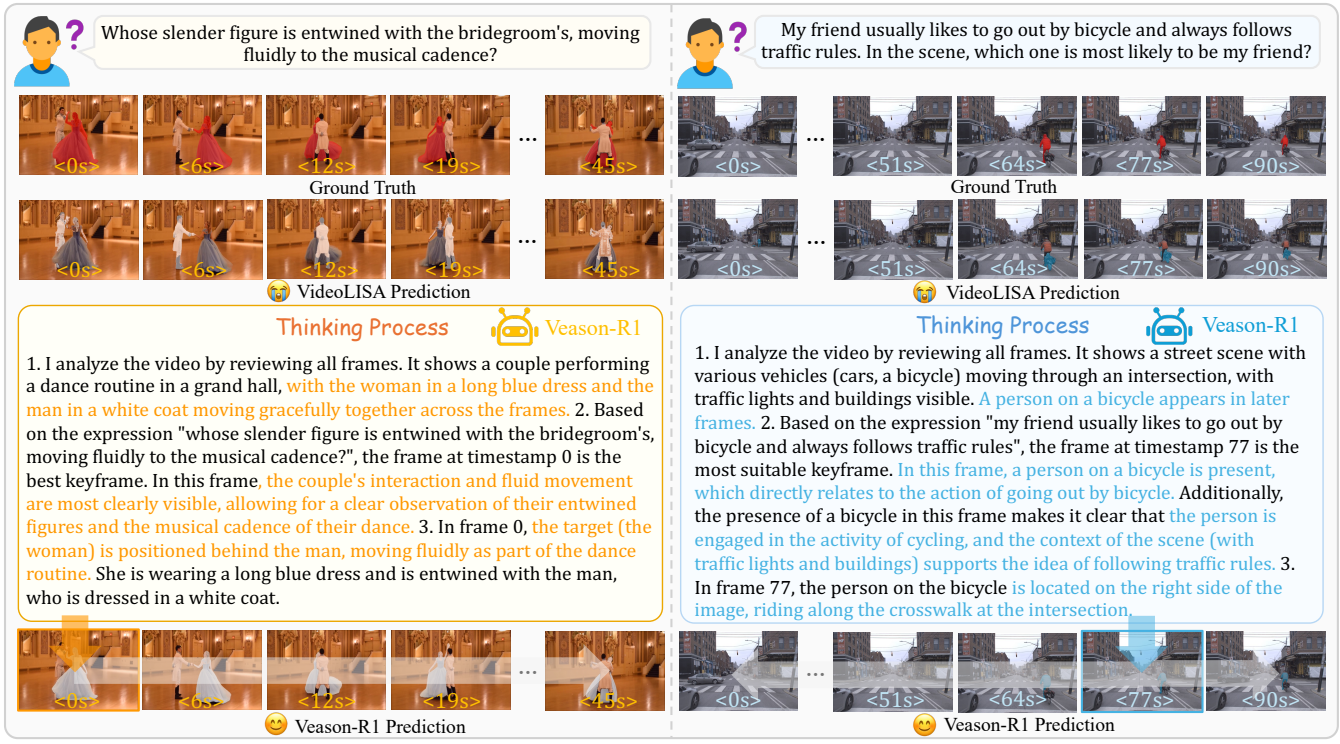


Figure 9: More visualization on the ReasonVOS dataset.